类别条件生成对抗网络的语音对抗样本生成方法

于振华，胡旭飞，叶鸥
（西安科技大学计算机科学与技术学院，710054，西安）

摘要：针对现有面向自动语音识别系统的对抗攻击方法难以捕捉不同语音尺度之间的相关性、导致攻击成功率低的问题，提出一种类别条件生成对抗网络的语音对抗样本生成方法。通过目标标签映射模块，将攻击目标标签转化为独热向量，作为条件输入到构建的类别条件生成对抗网络中，以此控制语音样本类别的生成。该类别条件生成对抗网络中的生成器，采用设计的 NReSidual U-block 网络模块与 U-Net 相融合，可以更好地学习不同时间尺度的语音特征，以及提升语音特征的表示能力，从而可以针对特定语音类别生成对抗样本；判别器采用卷积块和全连接层相结合的网络结构，将错误损失通过梯度反向传播至生成器，能有效保留语音信号的时序信息，并解决数据分布不稳定问题。在通用的谷歌命令数据集和音乐流派数据集上进行实验，结果表明，所提语音对抗样本生成方法的攻击成功率与主流方法相比，分别提高了 3.47%、5.1%，平均信噪比提升了3.2、1.49 dB，该方法具有较好的攻击效果和语音质量。

关键词：自动语音识别系统；生成对抗网络；对抗攻击；语音对抗样本生成；标签映射
中图分类号：TP399 **文献标志码**：A
DOI：10.7652/xjtuxb202412015 **文章编号**：0253-987X(2024)12-0153-12

Speech Adversarial Sample Generation Method Based on Class-Conditional Generative Adversarial Networks

YU Zhenhua, HU Xufei, YE Ou
(College of Computer Science & Technology, Xi'an University of Science and Technology, Xi'an 710054, China)

Abstract: To address the challenge posed by existing adversarial attack methods for automatic speech recognition systems, which struggle to capture the correlation between different speech scales, resulting in a low attack success rate, a novel speech adversarial sample generation method based on class-conditional adversarial networks is introduced. Through the target label mapping module, the target label is transformed into a one-hot vector, serving as a conditional input to the constructed class-conditional generative adversarial network to control the generation of speech sample categories. The generator in this network combines the designed NReSidual U-block network module with U-Net to better learn speech features at different time scales and enhance the representation capability of speech features, thereby enabling the generation of adversarial samples tailored to specific speech categories. The discriminator adopts a network structure combining convolutional blocks and fully connected layers, enabling the propagation of error loss back to the generator through gradient backpropagation to effectively retain the

收稿日期：2024-06-22。 作者简介：于振华(1977—)，男，教授，博士生导师。 基金项目：国家自然科学基金资助项目(62273272)。
网络出版时间：2024-08-29 网络出版地址：https://link.cnki.net/urlid/61.1069.T.20240828.1130.002

temporal information of speech signals and address data distribution instability issues. Experiments conducted on the Google command dataset and music genre dataset demonstrate that the proposed speech adversarial sample generation method increases the attack success rate by 3.47% and 5.1%, respectively, compared to mainstream methods. Additionally, the average signal-to-noise ratio is improved by 3.2 and 1.49 dB. This method exhibits good attack effectiveness and speech quality.

Keywords: automatic speech recognition system; generative adversarial network; adversarial attack; speech adversarial sample; label mapping

随着人工智能技术的发展，深度学习在众多研究领域已经取得一定成功，如语音识别[1]、计算机视觉[2]、自然语言处理[3]等，这些技术已经广泛应用于人们生活中。在语音识别领域中，自动语音识别系统(ASR)在无人驾驶、智能家居、语音助手等方面已经取得了显著进展，很大程度上归功于其内部所使用的深度神经网络。随着语音识别研究的不断深入，基于深度神经网络的语音识别模型识别率已大幅提高，但其安全性和鲁棒性的漏洞也随之暴露出来[4]。

文献[5]指出，将精心设计的微小噪声或扰动添加到原始语音中，就能使得自动语音识别系统产生误识别，并可能被恶意攻击者利用，这种含有噪声或扰动的语音被称为语音对抗样本。例如，恶意攻击者利用构建的语音对抗样本对自动驾驶汽车的ASR进行攻击，导致系统将原本用于触发刹车操作的语音指令错误地识别为执行加速操作指令。这种识别错误会导致自动驾驶汽车突然加速，从而引发严重的交通事故。由此可见，语音对抗样本对ASR的可靠性和安全性会造成较大影响，对语音识别应用安全构成严重威胁[6]。为减少ASR潜在安全问题，研究人员通过语音对抗样本检测系统漏洞，以此提升ASR的安全性和稳定性[7]。因此，研究ASR对抗攻击方法具有较大的理论意义和实用价值。

近年来，语音对抗样本生成方法研究备受学术界和工业界关注，如何设计并生成具有高攻击率与高质量的语音对抗样本仍然是一项具有挑战性的任务。现有的语音对抗攻击方法，例如基于梯度符号[8]的对抗样本生成方法，通过梯度信息直接优化输入，可以快速找到可行的对抗样本，但其通常需要多次迭代以达到理想攻击效果，生成对抗样本质量和攻击成功率受限于梯度信息的准确性，并且梯度信息无法准确地反映多尺度语音特征；基于遗传算法[9]的生成方法不依赖梯度信息，在全局优化问题上具有较强搜索能力，但收敛速度较慢，提取多尺度语音特征实现过程复杂；基于深度学习[10]的语音对抗样本生成方法利用神经网络自动学习语音特征，由于它的多层结构能够学习不同尺度的抽象表示，具有较强的多尺度特征提取能力。

Vaidya等[11]最早提出了对ASR的对抗攻击方法，其通过不断微调音频的梅尔倒谱系数语音特征，再将对抗性的音频特征重构为语音波形以生成对抗样本。Cisse等[12]通过在频域中添加扰动生成语音对抗样本，可以有效攻击DeepSpeech-2[13]深度语音识别模型。Esmaeilpour等[14]提出了一种使用Cramér距离来度量语音对抗扰动的方法，同时使用基于最小平方重构的方法，提高准确性和可靠性。Gupta等[15]提出了基于遗传算法的语音对抗攻击方法，该方法通过生成大量候选样本，并计算适应度分数进行变异，实现了对ASR有目标的黑盒攻击。Taori等[16]将遗传算法和梯度估计相结合，在对抗扰动陷入局部最小值时加快收敛速度并增加突变概率，提出新的动量突变更新算法，对标准的遗传算法进行改进。Du等[17]提出一种基于粒子群优化(PSO)的算法，利用修改后的PSO算法找到粗粒度扰动，再使用欺骗梯度方法修正对抗扰动，生成的对抗样本在ResNet18模型[18]上的攻击成功率达到99.45%。

近几年的语音对抗样本生成方法研究中，Carlini等[19]提出了一种基于迭代优化的方法，通过对原始语音波形进行优化并生成对抗样本。所提方法能够确保整个ASR系统实现快速收敛，该方法对DeepSpeech[20]的攻击成功率达到了100%。Neekhara等[21]在每个样本的初始扰动上迭代地添加扰动，最终找到适用于所有样本的通用对抗扰动，生成的样本对DeepSpeech的攻击成功率可达89.06%。Yuan等[22]使用可逆的MFCC提取模块，实现对原语音信号波形的修改，不断执行梯度下降来生成具有最小化扰动的语音对抗样本，以保证对用户的隐蔽性。Xie等[23]使用定长通用噪声

的重复回放，适应不同长度的语音输入，进而构造通用扰动。Kong 等[24]提出了一种基于生成对抗网络(GAN)的语音对抗样本生成方法，使用基于 Wasserstein 距离的损失函数，提高生成对抗样本的质量，通过在原始语音信号中添加对抗噪声作为生成器网络的输入，生成具有误导性的对抗样本。以上方法一方面受到数据规模和数据质量的限制，无法充分捕捉到语音信号中复杂的多尺度信息特征；另一方面容易受到微小干扰的影响，使得生成的对抗样本往往无法准确控制其最终的语音特征，导致无法控制最终生成样本的趋势。条件生成对抗网络(CGAN)[25]可以在一定程度解决上述问题，通过合理的选择条件信息和网络结构来改善语音对抗样本最终生成的结果。

因此，本文提出一种类别条件生成对抗网络(CC-GAN)，该模型设计了 NRSU 网络结构与生成器的 U-Net 不同层网络相融合，以便更好地学习不同时间尺度下的语音特征，并提高其对语音特征表达能力，从而生成高攻击率和高质量的语音对抗样本。

本文设计了语音类别标签映射模块，该模块将语音信号转换为类别标签，并将这些标签的标量形式映射为类别特征向量。通过将这些特征向量作为条件信息输入到生成对抗网络中，指导生成器生成符合特定语音类别的对抗样本。在谷歌语音命令数据集 Speech Commands[26]和音乐流派数据集 GTZAN[27]上进行实验验证，并与现有生成方法进行对比，本文方法生成的对抗样本攻击成功率与语音质量均较高，为指导更有效的语音对抗样本生成提供了理论依据。

1　类别条件生成对抗网络

本文提出了一种基于类别条件生成对抗网络的语音对抗样本生成方法，其核心在于输入语音样本和特定类别信息，以使生成器产生对应类别的对抗扰动，如图 1 所示。

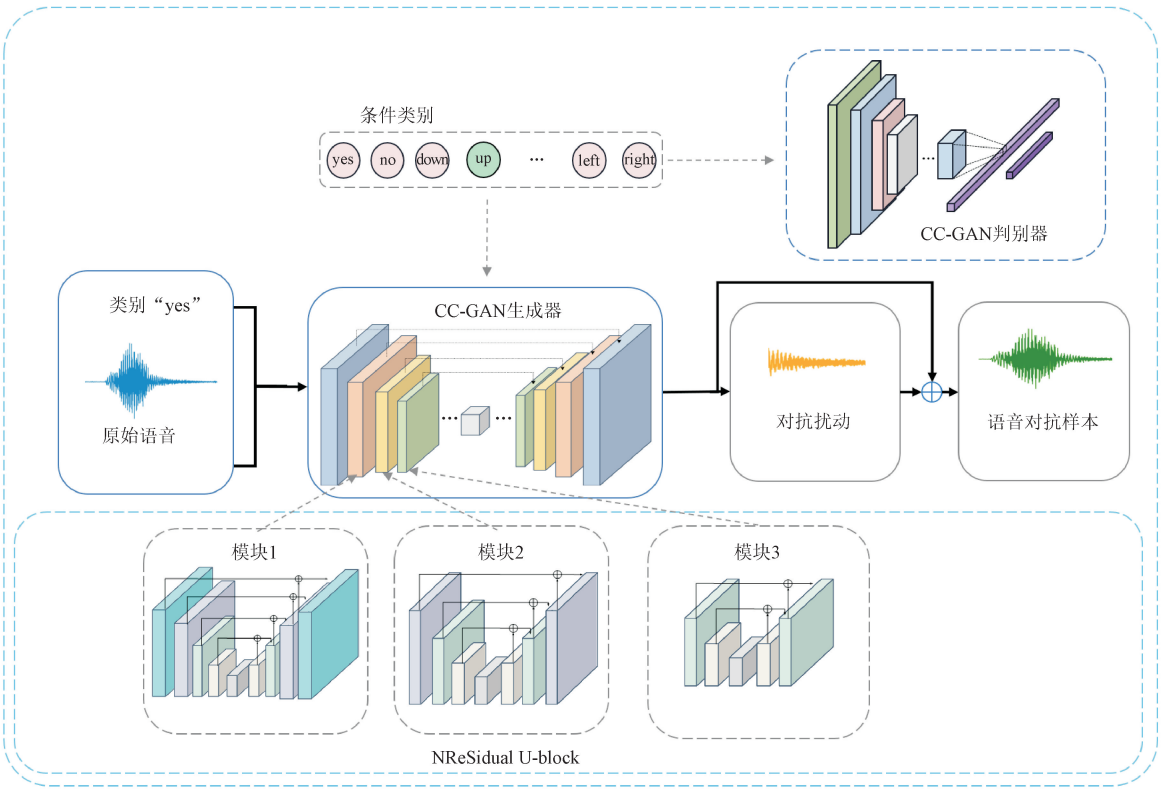


图 1　类别条件生成对抗网络整体框架

Fig. 1　The overall framework of the category conditional generative adversarial network

1.1　目标标签映射模块

为了使 CC-GAN 的生成器能够生成不同类别的语音对抗样本，本文设计了目标标签映射模块。这个模块将目标类别转换为目标类别向量，并作为条件信息输入到 CC-GAN 中。由于类别标签是标量数据，而语音样本通常是一维向量，可能导致目标

类别信息被忽视。本文设计的目标标签映射模块将标量数据转换为高维特征向量，对于特定的目标类别，映射后的特征向量可以表示为

$$\boldsymbol{t} = Z\boldsymbol{e}_y \tag{1}$$

式中：$\boldsymbol{e}_y$ 为用于编码目标类别 $\boldsymbol{y}$ 的独热向量，独热向量的维度等于类别的数量。假设有 10 个类别，则每个类别可以表示为一个 10 维的独热向量。在独热向量中，只有对应类别的维度为 1，其余维度均为 0，其位置表示目标类别的索引。Z 为线性变换参数，其特征向量维度是生成器输入的维度。独热向量 $\boldsymbol{e}_y$ 和线性变换参数 Z 相乘，得到一个高维向量 $\boldsymbol{t}$，将高维向量 $\boldsymbol{t}$ 与语音样本进行拼接构成生成器的输入。目标类别映射模块将目标类别转换为独热向量，然后使用 Linear 函数定义全连接层并进行前向传播。

1.2 类别条件生成对抗网络生成器

CGAN 的核心思想是通过生成器网络和判别器网络进行对抗学习，最小化生成样本与真实样本之间的差异来提高生成器的性能，并允许通过给定条件信息来生成特定类型的样本。CGAN 目标函数如下

$$\min_G \max_D V(D,G) = E_{x\sim p_d(x),\boldsymbol{y}\sim p_{d(\boldsymbol{y})}}[\log D(x,\boldsymbol{y})] + E_{z\sim p_z(z),\boldsymbol{y}\sim p_{d(\boldsymbol{y})}}[\log(1-D(G(z,y),\boldsymbol{y}))] \tag{2}$$

式中：$E[\cdot]$ 为期望函数；x 为真实样本；z 为随机噪声；$\boldsymbol{y}$ 为目标类别；$p_d(\boldsymbol{y})$ 为概率函数，生成器 G 通过随机噪声 z 和类别信息 $\boldsymbol{y}$ 生成样本。式(2)的目标是使生成样本的分布无限趋近于真实样本分布，从而误导判别器 D 将生成的样本分类为真实样本。直接对目标函数进行求解是非常困难的，因此在训练过程中一般采用固定一个优化目标的策略，如先固定生成器再训练判别器。

受 CGAN 的启发，本文生成器目标函数包含两部分：①对抗损失，即通过与判别器的对抗训练，最小化生成样本与真实样本之间的差异；②类别一致性损失，即为了确保生成样本的类别信息与目标类别一致，在目标函数中采用交叉熵损失函数来计算生成样本的类别标签与目标类别标签之间的差异。该目标函数如下

$$\min_G \max_D V(D,G) = E_{x\sim p_d(r),\boldsymbol{y}\sim p_d(\boldsymbol{y})}[\log D(r,\boldsymbol{y})] + E_{r\sim p_d(z),\boldsymbol{y}\sim p_d}(\boldsymbol{y})[\log(1-D(G(z,Z\boldsymbol{e}_y),\boldsymbol{y}))] \tag{3}$$

式中：r 为原始语音样本；目标类别向量 $\boldsymbol{y}$ 为条件输入；$Z\boldsymbol{e}_y$ 为映射后的特征向量；生成器 G 根据噪声 z 和映射后的特征向量 $Z\boldsymbol{e}_y$ 生成对抗样本。本文的目标函数旨在通过对抗学习过程中优化生成器和判别器，使生成器能够在给定条件下生成尽可能逼真的样本，同时使判别器能够准确区分真实样本和生成样本。

本文的生成器采用编码器-解码器式的网络结构，该结构已被广泛应用于图像处理、语音处理和自然语言处理等多个领域的研究。生成器包含编码层和解码层：编码部分负责接收语音输入数据并通过一系列卷积层进行特征提取，这些卷积层逐渐减小特征表示的尺寸，同时增加通道数，捕捉语音输入数据的抽象特征；解码部分则将这些高维表示映射回原始数据空间，解码层使用 Tanh 激活函数，将解码层输出的特征向量映射到(−1,1)范围内。在输入语音样本与类别特征向量基础上，生成对抗扰动任务的难度小于由噪声直接生成自然语音，因此在编解码层内未使用激活函数。

本文生成器的网络结构由 8 个一维卷积层和 8 个反卷积层构成。其中，每个卷积层的内核大小为 1×32，步长为 2，包含多个卷积滤波器，每个卷积层的输出经过 PReLU 激活函数处理。这种网络结构可以将输入语音样本映射到表示样本特征的低维空间，通过反卷积层将低维特征表示映射成与输入样本尺寸相同的对抗扰动。反卷积层进行上采样操作，逐渐增加特征表示的尺寸，同时减少通道数。为防止训练过程中出现梯度爆炸或消失问题，本文在生成器中引入了跳跃连接。通过将编码器生成的特征图与解码器的相应反卷积层连接，实现卷积层与反卷积层特征的共享，避免扰动值范围过大，从而在生成对抗样本时保持良好的语音质量。

本文使用的语音采样率为 16 kHz，为了使其符合生成器的网络输入尺寸，在语音采样值后以补 0 的方式将长度为 16 000 的语音样本扩充长度为 16 384。将调整后的语音与目标标签映射向量连接，构造成大小为 2×16 384 的向量输入到生成器 G 中，生成对抗样本可表示为

$$r' = r + G(r,\boldsymbol{y}) \tag{4}$$

对抗样本 r' 被送入判别器 D，功能是区分对抗样本和正常样本，其目的是驱动生成器 G 学习并生成令判别器无法区分的对抗扰动。在生成器与判别器的对抗性训练中，需要找到一个最优的扰动，通过优化模型参数和对抗扰动，使得模型能够在存在对抗性扰动的情况下仍然保持高准确率。最优扰动的表达式为

$$c^* = \arg\min c(l(f(r+\boldsymbol{c}),\boldsymbol{y}) + \mu\max(0,\|\boldsymbol{c}\|_p - \delta)) \tag{5}$$

式中：$\boldsymbol{c}$ 为对抗扰动；μ 为惩罚系数；δ 为阈值。μ 用于控制扰动 c 的 p 范数约束条件强度，当 μ 较大时，超出阈值 δ 的扰动将导致目标函数值显著增加，从而更容易找到满足约束条件的解。此优化问题的目标是找到一个扰动 $\boldsymbol{c}$，使得扰动 $\boldsymbol{c}$ 的 p 范数小于给定阈值，并且令网络在扰动后的输入上的预测与目标 $\boldsymbol{y}$ 之间差距尽可能小。

结合生成器的 U-Net 结构，设计了 NReSidual U-block 网络，并将该网络与 U-Net 网络相融合，以此提升生成器对语音特征的提取能力。NRSU 模块用作构建嵌套 U 型结构的基本单元，生成器中 U-Net 网络的每个编码和解码阶段由设计的 NRSU 模块构成，通过这种方式，编码器能够从语音信号中提取特征，且解码器可以通过反卷积进行上采样特征。NRSU 模块包含输入卷积层、U 形结构及残差连接层 3 个关键部分。输入卷积层负责实现通道的转换，而 U 形架构则负责提取不同尺度的上、下文信息，通过残差连接层，实现输入层与中间层的有效连接。

在模块 1 的残差结构设计里，编码结构位于左边，采用带降采样的卷积层来执行连续的 3×3 卷积，以此提取语音的多层次特征并扩展其感受范围。解码部分设置于右侧，使用卷积层结合上采样操作进行数据恢复。在生成器中，前 3 层使用模块 1，第 4、5 层使用模块 2，最后两层使用模块 3。通过本文的实验验证，在生成器网络结构的第 3 层时，继续使用模块 1 会导致过多的下采样，从而丢失上下文信息，并出现过拟合现象。因此，在第 3 层之后，本文分别减少了上采样和下采样的次数，设计了相应的上采样和下采样模块，具体网络结构详细设计如表 1 所示。

表 1　NRSU 网络详细结构参数

Table 1 NRSU detailed structural parameters

输入数据	输出数据	输入通道	输出通道
(3，16 384)	(3，16 384)	3	3
(3，16 384)	(12，16 384)	3	12
(12，4 096)	(12，4 096)	12	12
(12，4 096)	(12，1 024)	12	12
(12，4 096)	(12，1 024)	12	12
(12，2 048)	(12，2 048)	24	12
(12，4 096)	(12，4 096)	24	12
(12，8 192)	(3，16 384)	24	3

NRSU 采用编码-解码网络结构，其中编码器采用卷积层构建，主要用于捕获语音中的局部特征。NRSU 获取语音特征表示的过程如下

$$X_l=\begin{cases}C_l^{(n_c,k_c,s_c,p_c)}(F_l(X_{l-1})), & l<L\\ D_l^{(n_d,k_d,s_d,p_d)}(F_l(X_{l+1}), & \\ C_l^{(n_c,k_c,s_c,p_c)}(F_l(X_{l-1}))), & l\geqslant L\end{cases}\tag{6}$$

式中：X_l 为第 l 层的特征表示；C_l 为编码器的卷积操作；n_c 为编码器的卷积核数量；k_c 为编码器的卷积核大小；s_c 为编码器的卷积步长；p_c 为编码器的填充大小；D_l 为解码器的卷积操作；n_d 为解码器的卷积核数量；k_d 为解码器的卷积核大小；s_d 为解码器的卷积步长；p_d 为解码器的填充大小；F_l 为 NRSU 单元的特征表示；L 为生成器中 U-Net 的层数。

NRSU 获取语音特征表示的过程根据层数可以分为编码阶段和解码阶段。编码阶段中，将 NRSU 单元特征表示作为输入传递给编码器的卷积操作，编码器的卷积操作使用参数对特征表示进行处理，得到下一层的特征表示 X_l；在解码阶段，首先将 NRSU 单元特征表示 作为输入传递给解码器的卷积操作，同时将编码阶段得到的特征表示与解码器的卷积操作结果结合。解码器的卷积操作使用 n_d、k_d、s_d、p_d 对特征表示进行处理，最终得到下一层的特征表示 X_l。

NRSU 的编码层包含多个卷积层，每个卷积层具有不同的核大小、步长和填充方式，旨在捕获不同尺度的语音特征。在每个卷积层之后，使用批量归一化(BN)和激活函数来增强网络的非线性表示能力。解码层用于还原编码器捕获的特征以生成最终结果，其结构类似于编码器，但使用上采样操作逐步恢复语音信号特征。在解码层中的每个层级，都将其对应编码层的特征与当前解码层的特征进行融合，以增强多尺度特征的表示能力。本文在解码层部分引入了残差连接，可以有效缓解梯度消失问题，提高模型训练速度和性能，从而生成更具攻击性、且质量更佳的语音样本。生成器中每层卷积操作的计算复杂度为 $O(L\times k\times C_{in}\times C_{out})$，其中 L 为语音样本长度，k 为卷积核尺寸，C_{in}、C_{out} 分别为输入、输出通道数。

1.3　类别条件生成对抗网络判别器

判别器主要用于区分原始语音样本与生成器生成的样本，将目标类别特征向量与语音样本拼接，输入判别器中，促使生成器生成与真实样本分布相似的样本。

为了提取语音样本中有效的特征信息，并将其映射到低维空间，首先使用 1×31 的卷积核在前 11 个卷积块中进行卷积操作，这种卷积核既能保留语

音信号的时序信息，又能减小卷积块的尺寸，从而降低模型的参数数量，使得模型更轻量化。在每个卷积块后加入 BN 层和 Leaky-ReLU 激活函数，这种结构能够解决卷积层数据分布不稳定问题，保证模型训练的稳定性，并避免死神经元的出现。在全连接层之后，采用 softmax 函数将输出映射到[0，1]内，得到分类概率，使得判别器能够对语音样本的类别进行更加准确评估，从而提高模型分类性能。本文判别器网络结构采用了卷积块和全连接层相结合的深度学习结构，能够有效提取语音样本中的特征信息，并将其映射到低维空间，具有高分类性能等优点。判别器的主要计算复杂度集中在卷积和全连接上，总复杂度为 $O(m\times L\times k\times C_{\text{in}}\times C_{\text{out}}+n\times d_{\text{in}}\times d_{\text{out}})$，其中 m 为判别器中的卷积层数，n 为全连接层数，d_{in} 为全连接层输入维度，d_{out} 为输出维度。

1.4 损失函数

为保证生成对抗样本的语音质量和攻击成功率，受文献[28]的启发，本文采用的生成器的损失函数 L_{g} 可以表示为

$$L_{\text{g}}=E_x\log(1-D(r',y))+\lambda E_x[l_{\text{ce}}(f(r'),y)]+\eta\|r'-r\|_2 \tag{7}$$

式中：y 为攻击指定的目标类别；l_{ce} 为交叉熵损失函数；r' 为对抗样本；λ、η 为权重参数。为了使生成器所生成的对抗扰动保持在较小范围，本文使用基于二范数的内容损失函数，对扰动范围进行约束，随着该函数值的减小，对抗样本与原始样本的差异将不断缩小。

为了解决生成样本分布与真实样本分布的不准确问题，本文采用的判别器将真实样本与假样本映射到一个常量空间，并将错误损失通过梯度反向传播至生成器，使训练过程更加稳定。判别器损失函数可表示为

$$L_{\text{D}}=\frac{1}{2}(E_x\log(D(r',y_t))+E_x\log(1-D(r,y))) \tag{8}$$

本文对生成语音对抗样本采用了两阶段训练策略，首先进行 N 轮次的训练，第二阶段引入目标语音分类网络，进行持续训练。在交替训练生成器和判别器的过程中，目标语音分类网络的参数保持固定，仅将预测误差反向传播到生成器，以指导其训练过程，目标语音识别网络自身参数不进行更新，训练过程如下。

输入：目标标签 $\boldsymbol{y}$，训练数据集，迭代次数 i，预训练轮次 N，最大迭代次数 M，最小批大小 m，超参数 λ、η。

输出：经过训练的生成器 G^*

开始

(1)初始化：初始化生成器 θ_{G} 和判别器 θ_{D} 的网络参数。

(2)for 迭代次数 $i<M$ do

(3)从训练数据集中随机选择 m 个样本

(4)通过 $r'=r+G(r)$ 生成 m 个对抗性示例

(5)使用反向传播梯度 $\nabla\theta_{\text{G}}L_{\text{G}}$ 更新生成器参数 θ_{G}

(6)end if

(7)使用反向传播梯度 $\nabla\theta_{\text{D}}L_{\text{D}}$ 更新判别器参数 θ_{D}

结束

2 实验与结果分析

2.1 实验设置

本文实验使用的环境如下：Ubuntu 16.04 系统，处理器使用 Intel Core i9-10900K 3.7 GHz CPU，显卡为 NVIDIA GeForce RTX 2080Ti GPU，编程语言为 Python3.7，深度学习平台框架为 Pytorch1.4.0。

2.2 评价指标

本文在谷歌语音命令数据集 Speech Commands 和音乐流派数据集 GTZAN 上，对所提方法进行验证和评估。谷歌命令数据集包含约 65 000 个不同的简短语音片段，这些片段包含 30 个常见的命令词汇，如“yes”、“no”、“up”、“down”等。GTZAN 音乐流派数据集包含 1 000 首 30 s 长的音频片段，分为 10 个音乐流派，每个流派包含 100 首歌曲，流派包括蓝调、古典、乡村和迪斯科等。在实验中，本文采用 80%、10%、10%的比例将两个数据集分为训练集、验证集和测试集。在输入到 CC-GAN 生成器之前，首先对数据集中的语音进行归一化操作，本文对语音命令和音乐片段，使用式(6)进行归一化处理，将语音样本的采样值归一化到[−1，1]之间，以减少输入数据的差异，从而提高模型训练效果，归一化公式如下

$$N(x)=\frac{2}{65\,535}(x-32\,767)+1 \tag{9}$$

式中：x 为语音片段。生成对抗样本后，利用逆归一化函数将语音采样值转换成原始范围，逆归一化公式如下

$$I(x)=(x-1)\frac{65\,535}{2}+32\,767 \tag{10}$$

在实验中，本文将随机种子生成设置为 1 024，学习率设置为 1×10^{-4}，训练最大轮数设置为 50，最小批处理大小为 64。

生成的对抗样本对分类网络进行对抗攻击的主要性能评估指标是攻击成功率可表示为

$$\beta=\frac{M}{T} \tag{11}$$

式中：M 为被错误分类的样本数；T 为测试的所有样本数。式(11)衡量构建的对抗样本，被分类网络分类成目标类别的比例、比值越高越好。为评估语音对抗样本质量，本文使用常用的音频质量评估指标，信噪比为

$$S(r,r')=10\lg\frac{p_r}{p_{r'}} \tag{12}$$

式中：p_r、$p_{r'}$ 分别为原始语音信号和扰动噪声的能量。信噪比是将原始信号与扰动噪声进行比较，信噪比越大，语音越干净。本文使用语音质量感知评估(PESQ)和短时客观可懂度(STOI)这两个指标来评估生成对抗样本的语音质量。PESQ 是一种客观评估指标，可用于测量语音质量，取值范围为 -0.5～4.5，数值越高表示语音质量越好；STOI 是一种衡量语音可懂度的指标，可以反映出语音信号中单词被正确理解的百分比，数值范围为 0～1，数值越高表示可懂度越高。

2.3　攻击实验

本实验假设攻击者有权访问目标语音分类网络的模型结构和参数，即白盒攻击环境，将构造的对抗样本输入到目标网络中，实现对语音分类网络的目标攻击并计算其攻击成功率。此攻击实验使用的 ASR 系统是 WideResNet 和 SampleCNN 的分类网络，针对不同的 ASR，需要调整对抗样本的生成策略，以适应特定的语音分类网络特性。

针对语音分类网络 WideResNet、SampleCNN 进行攻击实验，评估生成对抗样本对上述两种网络的攻击效果。WideResNet 是一种基于 ResNet 的语音分类网络，通常用于语音唤醒等应用，对其攻击的实验结果以混淆矩阵形式呈现。图 2 展示了在 Speech Commands 数据集上攻击 WideResNet 网络成功率结果，图 3 展示了在 GTZAN 数据集上攻击 SampleCNN 成功率结果。实验结果表明，攻击 WideResNet 语音分类网络时，攻击成功率普遍超过 90%，平均攻击成功率为 95.69%。对于生成的对抗样本，目标分类网络平均准确率仅为 4.31%，说明本文生成的对抗样本几乎可以完全误导 WideResNet 的语音识别网络。例如，在定向攻击“up”为目标“down”的情况下，本文生成听起来像“up”的对抗样本可以被 WideResNet 误分类为“down”的概率高达 95.99%，这可能会对一些基于 WideResNet 的 ASR 系统造成严重安全风险。

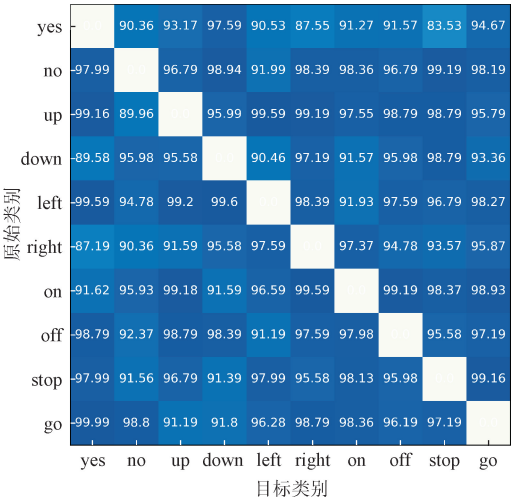


图 2　在 WideResNet 上的攻击成功率

Fig. 2　The attack_rate on WideResNet

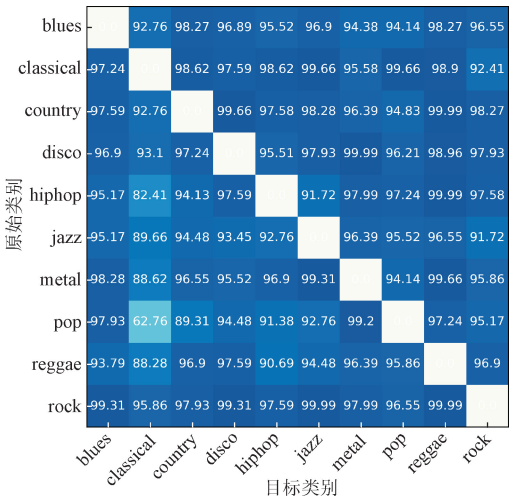


图 3　在 SampleCNN 上的攻击成功率

Fig. 3　The attack_rate on SampleCNN

在 WideResNet 网络中，针对“up”“right”“go”和“off”等常见单词，攻击成功率超过 95%。这是因为这些单词在语音识别系统中更为常见，相应的网络会更加倾向于将语音样本归类为这些单词，这也证明优化后的生成器具有较强攻击性能。对于攻击目标为“no”“left”“stop”等单词时，攻击成功率也分别达到 93.34%、94.69%和 95.76%。这是因为，这些单词可能在语音样本中存在一些特定的语音特征，容易被攻击者利用来生成对抗样本，分类网络也更容易将其误分类。攻击 SampleCNN 取得了较好的攻击效果，最大攻击成功率达到 99.99%，这表明

对于 SampleCNN，攻击者可以更容易地生成对抗样本，分类网络也更容易将其误分类。这是因为设计的 NRSU 网络可以在不同的尺度上提取和融合特征，使得每个嵌套的子网络可以在其对应的尺度上专注于特定的频域和时域特征。这种结构在生成对抗样本时，保持语音信号的细节和全局语义的一致性，减少生成过程中的信息丢失，因此在攻击过程中的可以保持较高的攻击成功率。

为了验证本文方法的有效性，与 3 个生成方法 GAN[28]、CM-GAN[29] 和 FGP-GAN[30] 进行比较。在相同的条件下进行两种攻击实验：在 Speech Commands 数据集上攻击 WideRestNet 模型，在 GTZAN 数据集上攻击 SampleCNN 模型。4 种方法在 WideResNet 和 SampleCNN 上攻击成功率如图 4、5 所示。实验结果表明，针对 WideResNet、SampleCNN，本文所提方法在攻击成功率方面优于其他 3 种方法，攻击分类网络的实验结果如表 2 所示。

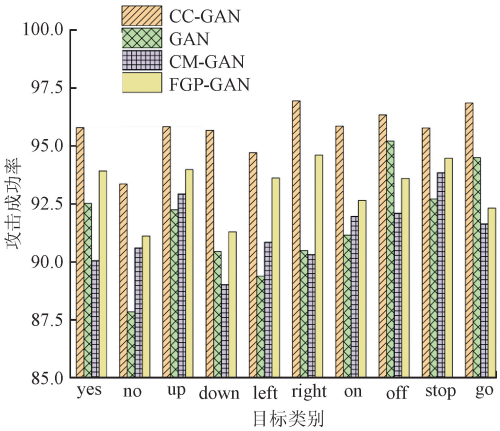


图 4 谷歌命令数据集上的攻击成功率

Fig. 4 The attack_rate on the Speech Command dataset

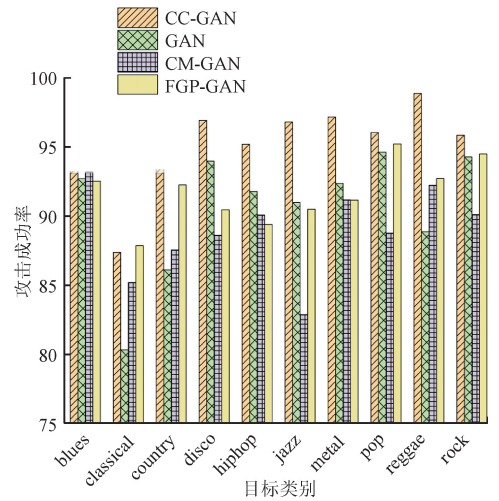


图 5 音乐流派数据集上的攻击成功率

Fig. 5 The attack_rate on the GTZAN

表 2 对比方法攻击成功率

Table 2 The attack_rate of comparison method

攻击方法	攻击成功率/%	
	WideResNet 模型	SampleCNN 模型
GAN	92.22	90.58
CM-GAN	90.65	88.96
FGP-GAN	93.14	91.64
CC-GAN	95.69	95.68

以 Speech Commands 数据集为例，在攻击 WideResNet 模型时，GAN 的平均攻击成功率为 92.22%，FGP-GAN 的平均攻击成功率为 93.14%，CM-GAN 的平均攻击成功率为 90.65%。相较于这 3 种方法，本文方法在攻击成功率方面达到 95.69%。对比方法 GAN，在 Speech commands 数据集上生成样本的攻击成功率提升了 3.47%，在 GTZAN 数据集上的攻击成功率提升了 5.1%。这是因为，NRSU 网络结构包含多个并行的卷积层，每个卷积层都具有不同的卷积核尺寸，使得 NRSU 能够捕捉到不同尺度和复杂程度的特征，在捕捉多尺度特征和学习更复杂的抽象表示方面具有更强的能力，从而显著提高了攻击成功率。

2.4 语音对抗样本质量评估

对原始语音样本进行扰动会导致对抗语音样本的质量出现一定程度下降，为了对生成的语音对抗样本进行质量评估，本文计算了所提方法与其他 3 种方法生成的语音对抗样本平均信噪比、PESQ 值和 STOI 值。表 3 展示了在 WideResNet、SampleCNN 两种数据集下，对比方法与本文方法生成对抗样本的语音质量评估结果。

CC-GAN 在信噪比和 PESQ 指标上均优于其他方法，尤其在攻击 SampleCNN 时效果最好，其 STOI 值最高，表明其生成的语音样本不仅清晰度高，而且可懂度更优。CC-GAN，在两种数据集中的所有指标上均显示出一致的优越性，说明其生成高质量对抗语音样本的任务中更为有效。对比方法 GAN，生成的样本在谷歌命令数据集上的信噪比提升了 3.2 dB，在音乐流派数据集上的信噪比提升了 1.49 dB。因此，CC-GAN 在生成高质量对抗语音样本方面表现更为出色。

CC-GAN 生成器在每个模块中使用了不同尺寸的卷积核来处理不同层次的特征，这种设计使得生成器能够更精细地调整各种特征的响应。这种精细的特征提取机制有助于生成器更好学习和模拟目标语音信号的细节特性。因此，生成的对抗样本不

仅质量更高,而且在保持原始语音特性的同时,不容易被轻易区分出来。CC-GAN 在生成对抗语音样本方面的隐蔽性,使得生成的样本几乎无法被人耳察觉。

表 3　不同方法产生对抗样本的语音质量

Table 3 The speech quality for generating adversarial samples using different methods

攻击方法	攻击 WideResNet			攻击 SampleCNN		
	信噪比/dB	PESQ 值	STOI 值	信噪比/dB	PESQ 值	STOI 值
GAN	20.27	3.087	0.952 4	26.92	3.947	0.957 9
CM-GAN	18.76	2.751	0.947 8	21.19	3.413	0.942 2
FGP-GAN	20.36	2.837	0.961 3	21.86	3.648	0.968 3
CC-GAN	23.47	3.136	0.972 4	28.41	3.991	0.974 7

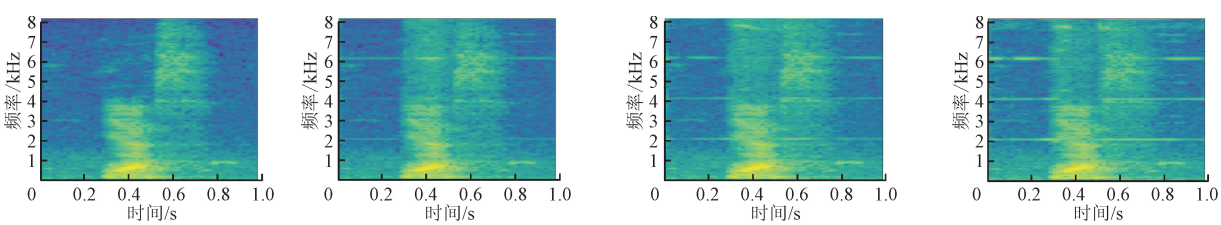
在两个数据集上生成对抗样本并对不同类别进行语音质量评估的实验中,结果如表 4 所示。本文生成器采用了卷积、转置卷积结合的方式进行上采样和下采样,可以在卷积进行特征压缩的同时,有效地进行特征扩展,从而保持音频信号的完整性和连贯性,以此学习到更复杂、更丰富的语音特征表示,通过在不同层次上融合特征并优化学习过程,从而生成更小扰动。因此,本文提出的生成器优化方法在衡量语音质量的性能指标上优于其他方法。

表 4　本文方法生成的对抗样本质量评估

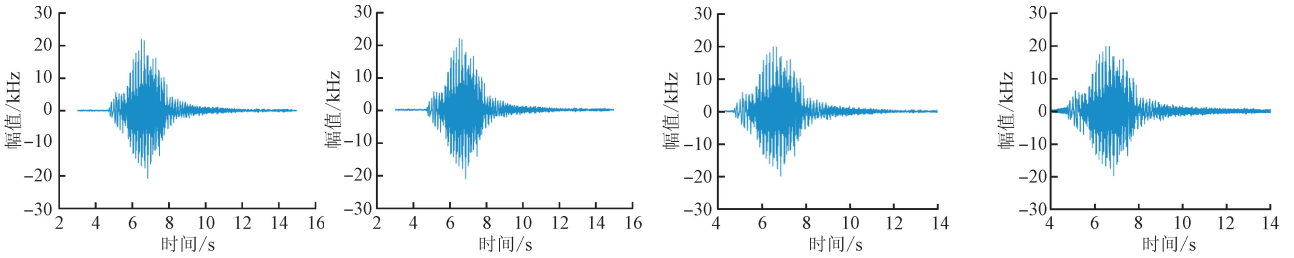
Table 4 Adversarial sample quality evaluation

类别	谷歌命令数据集										
	yes	no	up	down	left	right	on	off	stop	go	平均值
信噪比/dB	27.17	28.36	21.28	23.94	27.45	22.31	17.58	24.89	18.89	22.81	23.47
PESQ 值	3.429	3.072	2.813	2.916	3.397	3.374	2.981	3.246	3.219	2.914	3.136
STOI 值	0.966 1	0.978 6	0.998 7	0.999 4	0.961 5	0.989 1	0.864 1	0.976 2	0.994 0	0.996 5	0.972 4
类别	音乐数据集										
	blues	class	country	disco	hiphop	jazz	metal	pop	reggae	rock	平均值
信噪比/dB	29.14	27.21	31.46	28.13	28.49	26.98	25.81	33.68	24.73	28.50	28.41
PESQ 值	3.969	4.027	4.258	3.806	3.912	4.264	3.961	4.072	3.609	4.032	3.991
STOI 值	0.999 3	0.977 3	0.973 8	0.997 9	0.988 3	0.923 7	0.942 6	0.995 3	0.956 1	0.993 0	0.974 7

图 6 直观地比较了不同方法所生成的语音对抗样本,可以看出,本文对抗样本的波形与原始语音波形略微有些差异,CM-GAN、FGP-GAN 生成样本的波形与原始语音波形差异较大,而本文方法生成的样本则更为接近原始语音。实验结果表明,本文所提方法能够更好地保持音频质量和语音清晰度,生成更具攻击性和难以检测的对抗样本,相较于其他方法具有更高价值。为了更清晰地展示本文所提方法和其他方法的差异,本文计算并绘制了多种扰动波形图,如图 7 所示。由图 7 可以看出,本文方法生成的对抗扰动幅度小于对比方法,这表明生成器的特征融合策略在生成对抗样本时,能更好地捕捉和模拟原始音频信号中的细节特征,生成更具攻击性且难以检测的对抗扰动,以此生成具有较小扰动的对抗样本。因此,在生成的对抗扰动对比当中,本文生成的扰动相较于对比方法所来说具有更高的隐蔽性。



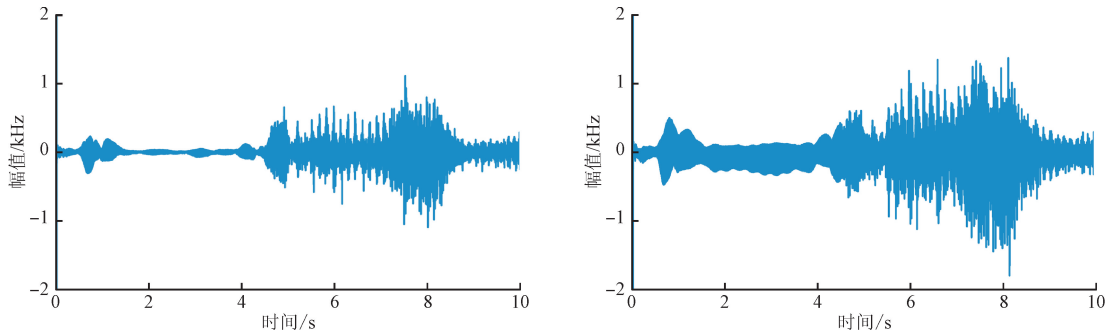
(a)原始语言语谱图　(b)CC-GAN 生成样本语谱图　(c)CM-GAN 生成样本语谱图　(d)FGP-GAN 生成样本语谱图



(e)原始语音波形图 (f)CC-GAN 生成样本波形图 (g)CM-GAN 生成样本波形图 (h)FGP-GAN 生成样本波形图

图 6 生成对抗样本的波形图及语谱图对比

Fig. 6 Comparison of waveform and spectrogram of generated adversarial sample



(a)CC-GAN 生成扰动波形图 (b)GAN 生成扰动波形图

图 7 生成对抗扰动的波形图对比

Fig. 7 Comparison of waveform plots for generating adversarial perturbations

3 结 论

本文提出了一种类别条件生成对抗网络的语音对抗样本生成方法。设计一个目标标签映射模块，将攻击目标标签转化为向量，作为类别条件生成对抗网络条件输入，然后改进生成器网络结构，通过为生成器设计的 NReSidual U-block 网络结构与 U-net 网络相融合，提升生成器对语音特征的提取能力。为给定语音片段和目标类别生成对抗扰动，对指定类别语音进行攻击。实验结果证明了本文所提方法可以生成攻击率高、扰动幅值更低和听觉感知质量更好的语音对抗样本。在未来的研究中，可以考虑不依赖于目标网络信息的黑盒攻击方法，并在攻击方法的设计中考虑物理世界的实际应用场景。

参考文献：

[1] LIU A H, HSU W N, AULI M, et al. Towards end-to-end unsupervised speech recognition [C]//2022 IEEE Spoken Language Technology Workshop. Piscataway, NJ, USA: IEEE, 2023: 221-228.

[2] ZHAO Xia, WANG Limin, ZHANG Yufei, et al. A review of convolutional neural networks in computer vision [J]. Artificial Intelligence Review, 2024, 57 (4): 99.

[3] LÜ Qing, APIDIANAKI M, CALLISON-BURCH C. Towards faithful model explanation in NLP: a survey [J]. Computational Linguistics, 2024, 50 (2): 657-723.

[4] SUN Zheng, ZHAO Jinxiao, GUO Feng, et al. CommanderUAP: a practical and transferable universal adversarial attacks on speech recognition models [J]. Cybersecurity, 2024, 7(1): 38.

[5] 于振华，殷正，叶鸥，等. 融合风格迁移的对抗样本生成方法 [J]. 西安交通大学学报，2024，58(7)：191-202.
YU Zhenhua, YIN Zheng, YE Ou, et al. Adversarial example generation method based on style transfer [J]. Journal of Xi'an Jiaotong University, 2024, 58 (7): 191-202.

[6] YUAN Xiaoyong, HE Pan, ZHU Qile, et al. Adversarial examples: attacks and defenses for deep learning [J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(9): 2805-2824.

[7] ESMAEILPOUR M, CARDINAL P, KOERICH A L. Cyclic defense GAN against speech adversarial attacks [J]. IEEE Signal Processing Letters, 2021, 28: 1769-1773.

[8] 邹军华，段晔鑫，潘雨，等. PIDI-FGSM：一种对抗样本生成的梯度处理新方法 [J]. 陆军工程大学学报，2022，1(5)：13-22.
ZOU Junhua，DUAN Yexin，PAN Yu，et al. Generating adversarial examples with PID iterative fast gradient sign method [J]. Journal of Army Engineering University of PLA，2022，1(5)：13-22.

[9] 李坤，郭威，张帆，等. 基于遗传算法的恶意软件对抗样本生成方法 [J]. 计算机科学，2023，50(7)：325-331.
LI Kun，GUO Wei，ZHANG Fan，et al. Adversarial malware generation method based on genetic algorithm [J]. Computer Science，2023，50(7)：325-331.

[10] KWON H，JEONG J. AdvU-net：generating adversarial example based on medical image and targeting U-net model [J]. Journal of Sensors，2022，2022(1)：4390413.

[11] VAIDYA T，ZHANG Yuankai，SHERR M，et al. Cocaine noodles：exploiting the gap between human and machine speech recognition [C]//Proceedings of the 9th USENIX Conference on Offensive Technologies. New York，USA：USENIX Association，2015：16.

[12] CISSE M，ADI Y，NEVEROVA N，et al. Houdini：fooling deep structured visual and speech recognition models with adversarial examples [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook，NY，USA：Curran Associates Inc.，2017：6980-6990.

[13] AMODEI D，ANANTHANARAYANAN S，ANUBHAI R，et al. Deep speech 2：end-to-end speech recognition in English and Mandarin [C]//Proceedings of the 33rd International Conference on International Conference on Machine Learning. Chia Laguna Resort，Sardinia，Italy：PMLR，2016：173-182.

[14] ESMAEILPOUR M，CARDINAL P，KOERICH A L. Towards robust speech-to-text adversarial attack [C]// IEEE International Conference on Acoustics，Speech and Signal Processing. Piscataway，NJ，USA：IEEE，2022：2869-2873.

[15] GUPTA H，WADHWA D S. Speech feature extraction and recognition using genetic algorithm [J]. International Journal of Emerging Technology and Advanced Engineering，2014，4(1)：363-369.

[16] TAORI R，KAMSETTY A，CHU B，et al. Targeted adversarial examples for black box audio systems [C]//2019 IEEE Security and Privacy Workshops. Piscataway，NJ，USA：IEEE，2019：15-20.

[17] DU Tianyu，JI Shouling，LI Jinfeng，et al. SirenAttack：generating adversarial audio for end-to-end acoustic systems [C]//Proceedings of the 15th ACM Asia Conference on Computer and Communications Security. New York，USA：Association for Computing Machinery，2020：357-369.

[18] ZAGORUYKO S，KOMODAKIS N. Wide residual networks [C]//Proceedings of the British Machine Vision Conference. Dundee，UK：BMVA，2016：1-12.

[19] CARLINI N，WAGNER D. Audio adversarial examples：targeted attacks on speech-to-text [C]//2018 IEEE Security and Privacy Workshops. Piscataway，NJ，USA：IEEE，2018：1-7.

[20] NASSIF A B，SHAHIN I，ATTILI I，et al. Speech recognition using deep neural networks：a systematic review [J]. IEEE Access，2019，7：19143-19165.

[21] NEEKHARA P，HUSSAIN S，PANDEY P，et al. Universal adversarial perturbations for speech recognition systems [C]//Proceedings Interspeech 2019. Graz，Austria：ISCA，2019：481-485.

[22] YUAN Xuejing，CHEN Yuxuan，ZHAO Yue，et al. Commandersong：a systematic approach for practical adversarial voice recognition [C]//Proceedings of the 27th USENIX Conference on Security Symposium. New York，USA：USENIX Association，2018：49-64.

[23] XIE Yi，SHI Cong，LI Zhuohang，et al. Real-time，universal，and robust adversarial attacks against speaker recognition systems [C]// IEEE International Conference on Acoustics，Speech and Signal Processing. Piscataway，NJ，USA：IEEE，2020：1738-1742.

[24] KONG J，KIM J，BAE J. HiFi-GAN：generative adversarial networks for efficient and high fidelity speech synthesis [C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook，NY，USA：Curran Associates Inc.，2020：17022-17033.

[25] GAUTHIER J. Conditional generative adversarial nets for convolutional face generation [EB/OL]. [2024-05-06]. http://www. foldl. me/uploads/2015/conditional-gans-face-generation/paper. pdf.

[26] MEHRA S，RANGA V，AGARWAL R. Improving speech command recognition through decision-level fusion of deep filtered speech cues [J]. Signal Image and Video Processing，2024，18(2)：1365-1373.

[27] TZANETAKIS G，COOK P. Musical genre classification of audio signals [J]. IEEE Transactions on Speech and Audio Processing，2002，10(5)：293-302.

[28] WANG Donghua，DONG Li，WANG Rangding，et

al. Targeted speech adversarial example generation with generative adversarial network [J]. IEEE Access, 2020, 8: 124503-124513.

[29] ZHANG Wenbin, LIAO Jun, ZHANG Yi, et al. CMGAN: a generative adversarial network embedded with causal matrix [J]. Applied Intelligence, 2022, 52 (14): 16233-16245.

[30] LIU Xin, ZHANG Weiwei, ZHENG Zhaohui, et al. FGP-GAN: fine-grained perception integrated generative adversarial network for expressive mandarin singing voice synthesis [J/OL]. IEEE Transactions on Consumer Electronics, 2024. (2024-06-11)[2024-05-06]. https://doi.org/10.1109/TCE.2024.3412053.

（编辑 赵炜）